Automotive Campus 30
5708 JZ Helmond
P.O. Box 756
5700 AT Helmond
The Netherlands

www.tno.nl

T +31 88 866 57 29
F +31 88 866 88 62

TNO report

TNO 2020 R12162 | Final report

Human Factors Guidelines Report 5: Test Criteria

Date	19 Februari 2021
Author(s)	Jan Souman, Marijke van Weperen, Jeroen Hogema, Marika Hoedemaeker, Frank Westerhuis, Arjan Stuiver, Dick de Waard, Karel Brookhuis
Copy no	
No. of copies	
Number of pages	16
Number of appendices	
Sponsor	Rijkswaterstaat
Project name	RWS Human Factors Guidelines
Project number	060.45606

Contents

1 Introduction 3

2 Developing test criteria 5
2.1 From guidelines to evaluation criteria 5
2.2 Development of evaluation criteria 5
2.2.1 Step 1: Definition 6
2.2.2 Step 2: Operationalization 6
2.2.3 Step 3: Setting Test Criteria 7
2.2.4 Step 4: Selecting Evaluator(s) 8
2.3 Test methods 8
2.3.1 Standardization of test procedure 9
2.3.2 Test scenario selection 10
2.3.3 Test execution 11
2.4 Factors influencing test criteria 12
2.5 Knowledge gaps 14

3 Conclusions 15

4 References 16

1 Introduction

Driver assistance systems and automated vehicle systems will only be able to realize their full potential in terms of safety effects if they take the end-user into account in their design. In 2019, the Ministry of Infrastructure and Water Management commissioned “Human Factors guidelines for safe in-car traffic information services” [ID5308][1]. These guidelines are intended to provide both policy makers and manufacturers / service providers with guidance in the safety assessment of nomadic devices in vehicles, in particular devices that provide information, such as navigation systems.

In recent years, however, there has also been a strong increase in driver assistance systems, ADAS (Advanced Driver Assistance Systems), which interact with the driver, support tasks, and sometimes even (partly) take over the driving task. The current version of the guidelines contains little or no guidelines specifically related to ADAS. In view of the current developments, it is advisable to expand the guidelines with these types of systems, allowing both system designers and policy makers to take these into account. Here, we follow the definition of ADAS as given by the Dutch Safety Board: *“Advanced Driver Assistance Systems (ADAS) are systems that assist the driver in carrying out the primary driving task. ADAS observe the environment using sensors and are able to take over control of speed or driving direction, subject to the responsibility of the person at the wheel. Systems of this kind are also able to warn the driver in situations that the system considers dangerous.”* [ID14] Where possible, Automated Driving Systems (ADS) will also be included in the development of the HF Guidelines.

If there are guidelines that a design must meet, these guidelines can also be used to check if the design complies with them. In other words, where the "HF Guidelines" specify what should be taken into account in the design of in-vehicle systems, they can also be used for the evaluation of these systems when the guidelines are combined with evaluation tools and criteria. After all, a good system must comply with the guidelines. In the end the objective of the development of the "HF Guidelines" is to arrive at a uniform evaluation framework of the interaction processes between vehicle and driver.

RWS has asked Rijksuniversiteit Groningen (RUG) and TNO to provide these Human Factor Guidelines for ADAS and Automated Driving Systems.
To come to these guidelines a number of separate reports have been prepared:

- Report 1: Literature review and overview
- Report 2: Overview and description of the different driver support systems
- Report 3: Literature study on the use of ADAS and the mental models of drivers
- Report 4: Human Factor Guidelines for ADAS and Automated Driving Systems

[1] The ID numbers between square brackets refer to the ID in the repository as explained in Report 1 [ID5357].

- Report 5: Overview of required knowledge to convert HF guidelines into an evaluation tool.

Reports 1, 2, 3 and 4 have previously been written in this project. The current report (Report 5) describes the steps that are necessary to develop the Guidelines as described in Report 4 into an evaluation tool, which can be used to assess whether a given system adheres to these Guidelines.

2 Developing test criteria

2.1 From guidelines to evaluation criteria

As mentioned in Report 4, the HF guidelines for ADAS and ADS are design principles for safe and effective use of ADAS and ADS [ID5377]. In accordance, this should also mean that if a system design meets these principles, it should be a ‘good’ design that allows all drivers to use it safely and effectively. In turn, this implies that the guidelines can potentially be used as an evaluation tool as well. Even though this should be possible in theory, there are several issues that have to be addressed before they can be used for evaluation.

First, it should be noted that many of the original guidelines are aimed at guiding the design process and are not created for evaluating an end-product. For example, guidelines state what an ideal system should strive for, what it should incorporate, what designers should take into account, and how they can improve (parts of) a system. With some exceptions, however, many guidelines do not explicitly state when a system is acceptable or ‘good enough’, which is necessary information for generating evaluation criteria. Even though evaluative words such as ‘sufficient’, ‘timely’, ‘clear’, ‘safe’, ‘simple’, ‘minimal’, ‘accurate’, ‘appropriate’, and/or ‘adequate’ are used, these words are open to interpretation and therefore difficult to base an objective evaluation on. Furthermore, even though some guidelines provide cut-off criteria, these are not all-encompassing. Indeed, what may be ‘sufficient’ or ‘on time’ for one system, might not be applicable to another system, another user, or even be similar for the same system and user in a different situation. This also highlights a second issue: the large number of guidelines and system elements that could be evaluated. Indeed, in Report 4, 60 guidelines are provided which may all be used as input for evaluating (parts of) one system because the guidelines are very broadly defined [ID5377]. Therefore, the guidelines must first be prioritized and those most relevant for evaluation need to be selected.

2.2 Development of evaluation criteria

After the most relevant guidelines have been selected, these require further consideration in order to develop them into concrete ADAS evaluation criteria. In this paragraph, suggestions for steps towards creating these criteria will be provided. The following guideline from Report 4 [ID5377] will be used to illustrate the development process:

> *System state changes of assistance systems should be communicated timely and effectively.*

Developing test criteria involves three steps: definition of guideline elements, operationalization, and defining test criteria. In addition, it has to be decided who will conduct these tests.

2.2.1 Step 1: Definition

Even though the abovementioned guideline might seem straightforward, several elements require definition before an evaluation can be performed: ‘system’, ‘state’, ‘changes’, ‘timely’ and ‘effectively’ (see also Section 2.2.3). First, it should be defined what is meant by ‘system’. Indeed, in every vehicle, there are many systems available that provide some sort of information and even for this particular guideline, which is limited to ‘assistance systems’, chances are that there is more than one assistance system installed in a vehicle. As a guideline, this statement applies to all assistance systems, making it a useful tool for designers indeed. For evaluation purposes, however, it should first be determined which ‘assistance system’ is meant as this also determines which ‘states’ and which ‘changes’ can be evaluated. Even though most systems do share some interchangeable states (e.g. ‘on’, ‘off’, ‘active’, ‘stand-by), there are also states that may be system-specific that need additional specification.
Furthermore, a system may have many states or sub-states that are required to make the system function correctly, without each of these being directly relevant for the user.

By providing clear definitions, it should become clear for an evaluator what he or she should evaluate exactly, as the following (simplified) example illustrates:

> *Lane Centering (LC) state change from ‘active’ to ‘stand-by’ should be communicated timely and effectively.*

If step 1 is performed, however, it quickly becomes apparent that this procedure greatly increases the number of evaluation possibilities for each guideline. This further emphasizes the necessity for pre-selecting guidelines because otherwise the numbers become too large to be used practically.

2.2.2 Step 2: Operationalization

Even though the abovementioned definitions provide clarity about what to evaluate, they do not provide any information about how to evaluate. Therefore, the second step should be aimed at operationalization. Ideally, it should be possible to measure exactly whether a system adheres to each guideline. Measurement theory distinguishes four different types of measurement scales:

1. Nominal scale: values indicate categories without ordering (e.g., 0 = male and 1= female)
2. Ordinal scale: the order of values is meaningful, but the differences between them are not (e.g., 1 = totally disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = totally agree)
3. Interval scale: the differences between values are meaningful, but their ratios are not because the scale has an arbitrary zero point (e.g., temperature measured in degrees Celsius or Fahrenheit)
4. Ratio scale: ratios between values are meaningful, because the scale has an absolute zero point (e.g., temperature in degrees Kelvin or speed in km/h)

For evaluation purposes, measurements on at least an ordinal scale are necessary, in order to be able to compare systems in terms of their adherence to a guideline (A is better than B). However, they do not provide a clear cut off point, since the values in themselves are not meaningful. In addition, measurements on an ordinal scale are more likely to be subjective, since they always involve someone who has to make a judgment concerning which is better. Therefore, if possible, guidelines

should be operationalized to an interval or ratio scale, allowing for quantitative testing. This is also likely to be more objective and less dependent on the person who performs the test. Examples are measuring reaction time in seconds or milliseconds, time spent looking within a certain region of interest or the gaze angle relative to a point of reference. The challenge is to translate guideline terms such as ‘effective’ or ‘safe’ into such variables that can be measured quantitatively. Often, new measurement methods need to be developed.

Ideally, measurements should be independent of the person carrying out the measurements. In practice, this is not always easy. Especially for things that can only be measured on an ordinal scale, system evaluation may involve a subjective element. This can be mitigated by providing standardized test methods, such as questionnaires or rating scales. Moreover, measurement validity and reliability can be affected by characteristics of the person who carries out the measurements. Experts may be able to rate system behaviour more consistently than naïve users (higher reliability), but they may not rate it in the same way, for instance because they have better knowledge of how a system works. Therefore, their evaluation may not apply to other user groups (lower validity). Conversely, naïve users may be more variable in their judgments than experts (lower reliability), but better reflect the ratings by the average user (higher validity).

2.2.3 Step 3: Setting Test Criteria

After having developed and reached consensus about how and which user-system interaction aspects should be measure, it should also be determined which measurement values indicate that a system performs sufficiently. In other words: pass and fail criteria need to be developed. With regard to the guideline provided in Section 2.2.1, this means that the terms ‘timely’ and ‘effectively’ need concretization as well. Indeed, when is a warning given ‘on time' and when is it ‘effective’? To answer these questions, a threshold value (cut-off point) should be set based on the operationalization.

Finding appropriate evaluation criteria strongly depends on the available domain knowledge. For many elements of the guidelines, evidence-based information is only partially available. Empirical evidence that allows for setting clear test criteria is lacking in particular for the higher-level goals of the guidelines (e.g., correct interpretation and comprehension of information) and for relatively new systems. Even though some criteria are available for lower-level goals of the guidelines (e.g., display brightness, contrast, distance, viewing angle, text size, etc.), these criteria serve to increase the likelihood that a message is detected, but they do not guarantee that the content of that message is comprehended by the driver (interpretation). Especially for more complex systems (which most ADAS/ADS are), understanding of what the system does is an abstract process of which it is unclear when a driver knows enough (i.e., when the mental model is ‘sufficient’).

Returning to the example of communicating LC system state changes from active to stand-by in step 1. A typical situation in which the LC system deactivates itself is when the lane markings become invisible, due to bad weather or degraded lane markings. In this case, ‘timely’ should mean that the system informs the driver as fast as possible. However, at the same time, the system needs time to decide whether it should deactivate itself and inform the user, or that it can recover lane

marking information within a sufficiently short period of time. Therefore, ‘timely’ is determined not only by user response, but also by system capabilities and the situation. For illustration purposes, we will assume a maximum delay of 0.5 s. Therefore, the guideline should state that the system state should be communicated at most 0.5 s after line detection fails. This example would therefore lead to the following change of the guideline:

> *LC state change from ‘active’ to ‘stand-by’ should be communicated at most 0.5 seconds after the system fails to detect line markings.*

Thus, the very general guideline from the definition in 2.2.1 about timely and effective communicating the switch of ADAS from standby to active, has been specified in a few steps to the guideline above which is much closer to a guideline that can be evaluated (note that we did not make ‘effectively’ specific yet; this is even harder than ‘timely’). Note that this example does not imply that the user should also be able to respond within 0.5 s. It only concerns the expected system behaviour. How adherence to a guideline such as this can be evaluated is described in section 2.3 on test methods.

2.2.4 Step 4: Selecting Evaluator(s)

Evaluation of user-system interaction can either be used as consumer recommendations (c.f., safety tests by Euro NCAP[2]) or for type approval by vehicle authorities for new vehicle types (including road exemptions). In the second case, evaluation can be done by different parties. One obvious candidate is a governmental body such as vehicle authorities, while the responsibility for testing can also be put at vehicle manufacturers or ADAS suppliers. In the latter case, careful documentation of testing has to be provided to the relevant authorities for approval.

2.3 Test methods

The Human Factors Guidelines for ADAS and ADS are primarily intended for use by system designers and developers. A commonly used framework for system design and development is the so-called V-model. Figure 1 shows one of many variants of this model. Core elements are the decomposition of overall requirements into requirements and specifications of system components, the actual implementation or development of the system, and the verification and validation of system components as well as the complete system against the requirements. Guidelines mainly concern the definition and decomposition phases of system design. They formulate requirements and best practices for the system and its component from a certain perspective, in our case that of user-system interaction. However, they need to be complemented with testing in the integration and recomposition phases, to ensure adherence to the guidelines. Although the graphical representation of the V-model in Figure 1 suggests that these phases follow an orderly, linear path, reality is often more diffuse and involves multiple cycles through parts of this model.

[2] https://www.euroncap.com/en/ratings-rewards/

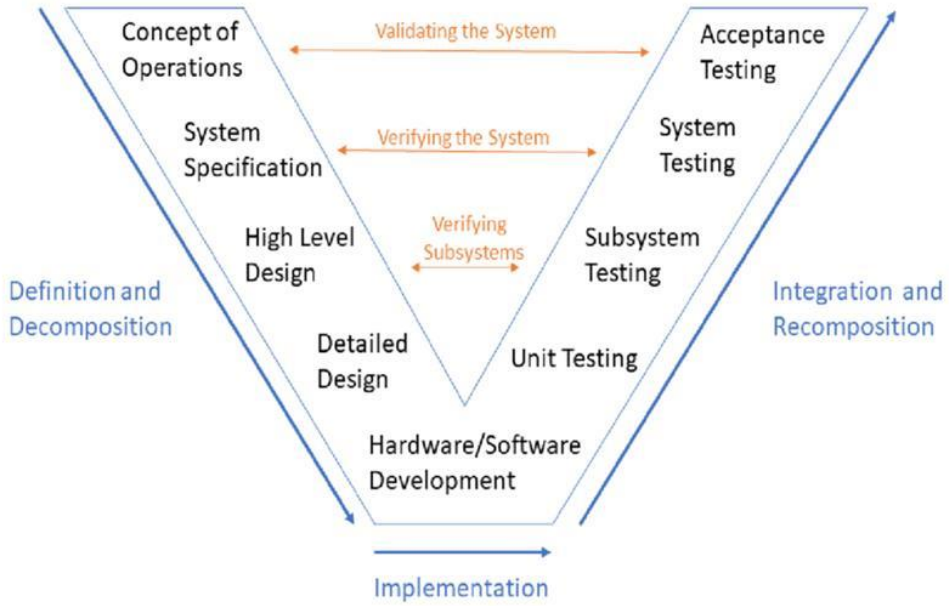

Figure 1. V-model of system design and development [ID5373].

Adherence to the guidelines should be tested by evaluating the total design and development process as described by the V-model and by assessing whether the Human Factors Guidelines have been followed throughout this process. Rather than just testing at the end of the development cycle whether a system meets the requirements that were derived from the guidelines, these should be guiding principles throughout the design and development process, ensuring that the user is taken into account at every step. Ideally, a commonly accepted process standard such as ISO 9241-210:2010 should be used for this. From a regulatory point of view, manufacturers can be required to document their adherence to such a standard throughout the design and development process and produce this documentation at market entry.

In addition, a test procedure at the end of the development process is needed to evaluate the end result in terms of usability in all its aspects. The question is how adherence to the guidelines can be tested throughout the design and development process, including the final evaluation. In addition, test moments after market entry may be necessary to evaluate system updates (which can be applied in the workshop with regular maintenance or directly Over The Air (OTA) through software changes). Below, we discuss several aspects of evaluation methods.

2.3.1 Standardization of test procedure

The test procedure should produce valid and reliable results. Therefore, the test criteria should be standardized as much as possible. The degree to which this is possible strongly depends on the steps described in Section 2.2 (definition, operationalization and test criteria). The most basic form involves the establishment

of checklists, which can be evaluated by trained experts. The most extensive form would be the definition of national or international standards.

1. Checklists:
As has been done for the previous Guidelines for Information Systems [ID5308], the Guidelines for ADAS and ADS can be translated into a simple checklist. To increase reliability, checklist-based evaluation can be conducted by trained experts. The difficult concern of deciding whether a system meets the guidelines falls onto the experts, which makes this method inherently subjective. Therefore, these experts must be trained to make their judgments more reliable and to reduce potential biases in the evaluation. To keep things workable, the checklist should not be too long, especially given the fact that multiple systems have to be evaluated, both separately and in combination. This requires prioritization and selection of the most relevant guidelines to be included in the checklist, possibly separately for different types of systems.

2. Standards:
Ideally, evaluation should take place according to a well-established and accepted standard, such as those from the ISO. The standard should describe not only the relevant guidelines, but also define their elements, prescribe how they must be tested and describe the test criteria for fail or pass. Compared to checklists, this makes the evaluation process far more objective, but also involves a far more elaborate process to develop. Test procedures should be defined to measure the relevant variables on at least an ordinal measurement scale, preferably on a ratio scale. As with checklists, standards will still require substantial prioritization and selection of the most important and relevant guidelines to use for system evaluation.

Either way, checklists or standards, defines the guidelines that are incorporated in the test, as well as how they must be tested.

2.3.2 Test scenario selection
In order to evaluate whether a given system conforms to the Guidelines, a selection needs to be made in which situations and scenarios the system needs to be tested. Testing a system in all possible situations is simply not possible, if only because this would take an infinite amount of time. Hence, relevant situations and scenarios have to be defined, prioritized and selected. Exemplary safety critical situations and situations that drivers most commonly experience should be selected as relevant. A database of relevant driving scenarios needs to be developed, which not only includes different assistance/automation systems, but also relevant driver types and states. A selection of these scenarios can then be used to evaluate ADAS or ADS on user-system interaction. Evaluation could be done using checklists or standards as described above. This approach is similar to the one used in Euro NCAP's crash tests, for example. From a large range of potential collisions, a few of the most common and/or dangerous ones are selected and tested in a standardized way. The biggest problem with this approach is to manage the combinatorial explosion of different scenarios that arises from combining all possible situations, systems, driver types and driver states. Just as problematic is covering scenarios that are relatively rare or those that have never happened because they involve new (combinations of) systems. Next to an empirical approach to scenario mining (i.e.,

collecting data in existing, not necessarily automated, vehicles), it may be necessary to include a more analytical approach, which predicts potentially problematic scenarios based on models of both systems and users. As with all scenario-based testing methods, care has to be taken that systems are not just optimized for a limited set of scenarios, loosing generalizability.

2.3.3 Test execution

Scenario-based testing can be implemented in two ways: user testing and simulation-based testing.

1. **User testing:**
   The most straightforward way to test user-system interaction is to evaluate systems in user trials. Users may be a small group of trained experts or large groups of untrained common users. Which of the two is most suitable depends on the specific system, scenarios and guidelines to be tested. Most likely, a combination of the two will be necessary. Experts are able to analyse system use in much more detail and with a larger scope than the average user. However, they may fail to be aware of the behaviour displayed by subgroups of users. For instance, testing user-system interaction in the case of driver fatigue requires testing with actual fatigued users. In general, expert testing will show less variation and therefore higher reliability than general user testing. However, this may come at the cost of validity, since experts may behave differently than (subgroups of) common users. Even when testing with naïve users, care has to be taken that the sample of test users is representative of the user population for which the results should apply (e.g., usability for elderly people should not be tested with students).

   User-based testing can be implemented in three ways:
   a. **Driving simulator tests:** User-system interaction can be evaluated in a driving simulator. One of the advantages of this approach is that it does not entail any safety risks for participants in terms of collisions and accidents. At the same time, the absence of these risks is also a drawback of this method, as it may change people's behaviour, thus affecting the absolute validity of the method (though research has shown that simulators can still have relative validity [ID5397]). Further advantages are that scenarios and situations can be created under experimental control, which is not that straightforward in the actual vehicle.
   b. **Test track tests:** As with driving simulators, test tracks allow for standardized testing procedures. Accident risks, though larger than in a simulator, can still be minimized by using dummies for other vehicles and road users. Participants' behaviour will most likely also be more similar to that in real traffic, compared to simulators.
   c. **On-road tests:** Testing user-system interaction in real traffic of course has the highest ecological validity. This comes at the cost of increased safety risks, especially in the case of new systems, complex situations and/or inexperienced users. Also, the unpredictable nature of traffic makes it much harder to follow standardized test procedures. Consequently, this method is most likely only suitable for a subset of systems, users and situations.

2. **Simulation-based testing:**
   User-system interaction can be evaluated in simulation, using models for both the user and the system. The advantages of this approach are that it can be completely standardized and automated, and does not entail any safety risks. In addition, model-based simulations can be used to evaluate envisioned systems that do not even exist yet, predicting their performance before actual development. However, this requires accurate models of both system and user. Though system models are often readily available, accurate user models are much harder to come by. They require modelling of the entire perception-cognition-action cycle, not only with respect to the driving task but also with respect to the interaction with the system under test. In addition, effects of driver state (fatigue, distraction) have to be modelled as well. Though several user models have been developed (e.g., [ID5374], [ID5375], [ID5376]), they generally concern limited parts of the driving task (e.g., car following behaviour) and are still far from being comprehensive and usable enough for the evaluation of user-system interaction.

2.4 Factors influencing test criteria

As already mentioned above, developing test criteria based on the Human Factors Guidelines for ADAS and ADS is a difficult endeavour. Several factors play a role.

- **Variation in systems:**
  The various systems themselves vary widely in their characteristics and user demands. Assistance systems require the user to play a different role than automation systems. Within assistance systems, warning systems such as BSW or FCW put different requirements on user-system interaction than active assistance systems such as ACC or LC, if only because they do not require constant monitoring of system performance like the latter do.
- **Combinations of systems:**
  Modern vehicles come equipped with an increasing number of assistance systems. As the behaviour of a specific system may depend on the presence and functionality of other systems, systems need not only be tested in isolation, but also, and particularly, in combination.
- **Variation in situations:**
  Different systems can be used in a huge range of situations, each putting different demands on user-system interaction. Parking systems are typically used in low speed / high workload scenarios, while for instance ADA is mostly used on highways with high speeds and varying workload. Some situations require immediate action upon system failure, such as when driving in tight curves or when crossing intersections, while others allow the user much more time to respond (for example driving on a quiet, straight highway). Therefore, for each system the most common situations as well as those implying the highest potential risks need to be defined and selected.
- **Variation in users:**
  Not only systems and situations vary, users do as well. Users differ in their characteristics and capabilities, because of experience, training, health, age or personality. Individual user state may also vary from day to day, depending on fatigue, the ability to concentrate, or otherwise. All of these

may impact the way users interact with ADAS and ADS and have to be taken into account when evaluating user-system interaction.

- **Variation in guideline characteristics:**
  Many of the guidelines described in Report 4 concern more abstract cognitive constructs, dealing not only with users' perception, but also with comprehension and decisions. In order to translate these guidelines into test criteria, they first need to be operationalized. This requires specifying these guidelines at a much more detailed and concrete level than their current state. It also implies that the current number of guidelines, which is already quite large, will grow even larger, as these abstract constructs may need to be operationalized into multiple measurable variables. This in turn again highlights the need for prioritization and selection of guidelines.
- **State of scientific knowledge:**
  Although there is a considerable body of scientific research concerning driving behaviour in general and the impact of ADAS/ADS in particular, there are still many unknowns. Some aspects, such as car following behaviour or how people steer their cars around bends, have been investigated and modelled extensively, but for others still a lot of research is needed in order to be able to decide which factors influence people's actions and which behaviour is adequate and appropriate. This is especially true for cognitive constructs such as mental models and situation awareness, which are thought to play an important role in driving and using ADAS/ADS, but have not been operationalized sufficiently to allow easy measurements. This makes it difficult to decide on clear test criteria for user-system interaction.
- **Technological developments:**
  ADAS/ADS technology is developing rapidly. New systems are introduced and existing systems are updated with new features, sometimes even literally overnight through over the air updates. This puts heavy demands on Human Factors Guidelines. On the one hand, they have to be general enough to apply to new systems and versions as well as existing systems. On the other hand, they have to be specific enough to be meaningful and useful. In addition, they should encourage safe use of new technology and not be an obstacle to new technological developments that could improve driving safety and comfort. These requirements apply to test criteria as well.

As described in Section 2.2, operationalization of the guidelines to ratio measurement scales allow for the objective and quantitative evaluation of user-system interaction. Some of the guidelines, in particular those dealing with the display and perception of information, contain constructs that are easily measured, such as the angle between display location and the driver's normal gaze position (e.g., Guideline 2.2.10), text and icon size (Guidelines 2.1.4 and 2.1.5), or number of colours used (Guideline 2.1.10). Others could be operationalized into similar quantitative measurements such as reaction time (e.g., Guideline 2.3.6). However, this will require a substantial amount of research, as the relationships between these variables and the underlying constructs are often complex and still poorly understood. Moreover, these measures are rarely both sensitive as well as specific for the constructs in question. This is especially the case for the more cognitive constructs referred to in the guidelines, such as over/underload of the driver (Guidelines 2.2.4 and 2.4.1), comprehension of information (Guideline 2.2.7),

predictability of system limitations (Guideline 2.3.1) or driver supervision of ADAS (Guideline 2.4.2).

2.5 Knowledge gaps

In the course of this project, we identified several areas where knowledge is insufficient to develop concrete and specific test criteria:

- Professional drivers (trucks, buses): very little is known concerning the interaction of professional drivers with ADAS and ADS. Even for platooning, a topic frequently researched in trucks, emphasis is more on the technical aspects than on user-system interaction.
- Individual differences: both differences between individuals (e.g., age, experience, training, personality traits) and within individuals (changes in driver state, caused by fatigue, sleepiness, alcohol or drug use, stress) are known to affect driving behaviour. However, little research has been devoted to how they affect the use of ADAS or ADS.
- User-accessible system descriptions: even though manufacturers advertise widely with the newest ADAS they have incorporated in their vehicles, it is very hard for common users to obtain information concerning how these systems work.
- Long term effects: most research is focused on users' first experiences with new ADAS. Very little research is available concerning long term effects in terms of actual use, impact on traffic safety and behavioural adaptation effects.
- Mental models: it is commonly accepted in traffic safety research that users' mental models of ADAS play an important role. However, it is far less clear how mental models develop or how they can be measured. This makes it difficult to quantify their impact on ADAS use.

These knowledge gaps are described here at a general level. Once the guidelines have to be translated into specific test criteria for specific systems in a specific situation for specific users, more concrete knowledge gaps will become apparent as well. However, these fall outside the scope of the current project.

3 Conclusions

In order to arrive at test criteria based on Human Factors Guidelines for ADAS and ADS, several steps need to be taken:

1. Selection of the most important and relevant guidelines, depending on the system that needs to be tested.
2. Each guideline needs to be translated into measurable variables and test criteria.
3. A set of test scenarios needs to be defined.
4. A suitable test method needs to be chosen (driving simulator, test track, on road, users).
5. Test procedures need to be defined in detail.

In practice, testing will need to be very selective, in terms of which systems are tested, which scenarios are used and which users are relevant. This implies that it is impossible to test all combinations of these comprehensively. Moreover, testing will often need to be tailored to the specific system in question.

In many cases, the available knowledge base is insufficient to make these selections and to translate the relevant constructs into measurable variables. In large part, this is because the literature only describes studies on existing systems. Therefore, scientific studies on new systems are by definition lacking. In addition, the knowledge that is available is often limited in its applicability and ecological validity, because it is mainly relevant to the specific systems and situations tested. Consequently, testing of new systems will have to be largely based on general principles and extrapolation from existing systems.

4 References

[ID14] Dutch Safety Board (2019). Who is in control? Road safety and automation in road traffic (report). The Hague, The Netherlands: Dutch Safety Board.

[ID5308] Kroon, E.C.M., Martens, M.H., Brookhuis, K.A., De Waard, D., Stuiver, A, Westerhuis, F., de Angelis, M., Hagenzieker, M.P., Alferdinck, J.W.A.M., Harms, I.M., Hof, T. (2019). Human factor guidelines for the design of safe in-car traffic information services (3rd edition). Groningen: University of Groningen.

[ID5351] Hogema, J., Souman, J., Stuiver, A., van Weperen, M., Westerhuis, F., de Waard, D. & Hoedemaeker, M. (2020). Repository Human Factors Guidelines driver assistance systems and automated vehicle systems (Repository in an MS Excel file.). Helmond: TNO.

[ID5354] Souman, J., van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A. & de Waard, D. (2020). Human Factors Guidelines report 2: Driver Support Systems Overview (TNO and RUG). TNO 2020 R12167.

[ID5355] Souman, J., van Weperen, M., Hogema, J., Hoedemaeker, M., Westerhuis, F., Stuiver, A. & de Waard, D. (2020). Human Factors Guidelines report 3: Use and Mental Models (TNO and RUG). TNO 2020 R12165.

[ID5357] Westerhuis, F., Stuiver, A., de Waard, D., Hogema, J., Souman, J., van Weperen, M. & Hoedemaeker, M. (2020). Human Factors Guidelines report 1: Literature review (TNO and RUG). TNO 2020 R12166.

[ID5374] Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. Psychological Review, 111(4), 1036–1060.

[ID5375] Salvucci, D. D. (2006). Modeling Driver Behavior in a Cognitive Architecture. Human Factors: The Journal of the Human Factors and Ergonomics Society, 48(2), 362–380.

[ID5376] Wiese, S., Lotz, A., & Russwinkel, N. (2019). SEEV-VM: ACT-R Visual Module based on SEEV theory. Proceedings of the 17th International Conference on Cognitive Modeling (ICCM 2019).

[ID5377] Souman, J., van Weperen, M., Hogema, H., Hoedemaeker, M., Westerhuis, F., Stuiver, A., de Waard, D. (2020). Human Factors Guidelines Report 4: Human Factors Guidelines for ADAS and ADS (TNO and RUG). TNO 2020 R12164.

[ID5378] Forsberg, K., & Mooz, H. (1991, October). The relationship of system engineering to the project cycle. In INCOSE International Symposium (Vol. 1, No. 1, pp. 57-65).

[ID5397] Kaptein, N. A., Theeuwes, J., & van der Horst, R. (1996). Driving Simulator Validity: Some Considerations. *Transportation Research Record: Journal of the Transportation Research Board*, 1550(1), 30–36.